机器学习导论
习题三

151250104, 卢以宁, kiwiloveskiwis@gmail.com

2017 年 4 月 21 日

1 [30pts] Decision Tree Analysis

决策树是一类常见的机器学习方法，但是在训练过程中会遇到一些问题。

(1) **[15pts]** 试证明对于不含冲突数据 (即特征向量完全相同但标记不同) 的训练集，必存在与训练集一致 (即训练误差为 0) 的决策树;

(2) **[15pts]** 试分析使用“最小训练误差”作为决策树划分选择的缺陷。

Solution. (1) 反设不存在训练误差为 0 的决策树，则必存在训练误差最小的决策树，设其值为 $\lambda > 0$. 则必存在测试样例 D_i 被划分错误。此时，设 D_i 所属叶子节点为 $Node_i$, 有如下两种情况:

1. $Node_i$ 仅包含 D_i 一个样例, 此时将 $Node_i$ 的类别标记改为 D_i 的类, 训练误差减小.

2. $Node_i$ 包含多于 D_i 的数个样例。由于各节点的特征向量各不相同, 我们可重复构造以 $Node_i$ 为根节点的子树, 重复划分直到存在某叶节点只包含 D_i 一个样例, 将其类别标记改为 D_i 所属类, 其余叶节点的标记维持不变。训练误差减小。

故与上述假设矛盾。故存在训练误差为 0 的决策树。

(2) 缺乏泛化性能，容易过拟合。例如, 在连续属性学习中，各个训练样例特征向量完全相同的概率很小, 此时我们可以按照 (1) 题的方式构造训练误差接近于 0 的决策树。但这样的坏处是过于针对训练数据，从而可能泛化误差更大。

2 [30pts] Training a Decision Tree

考虑下面的训练集：共计 6 个训练样本，每个训练样本有三个维度的特征属性和标记信息。详细信息如表1所示。

请通过训练集中的数据训练一棵决策树，要求通过“信息增益”(information gain) 为准则来选择划分属性。请参考书中图 4.4，给出详细的计算过程并画出最终的决策树。

表 1: 训练集信息

序号	特征 **A**	特征 **B**	特征 **C**	标记
1	0	1	1	0
2	1	1	1	0
3	0	0	0	0
4	1	1	0	1
5	0	1	0	1
6	1	0	1	1

Solution.

$$
\begin{aligned}
Ent(D) &= -\sum_{k=1}^{2} p_k \log_2 p_k = -(\frac{1}{2}\log_2 \frac{1}{2} \cdot 2) = 1 \\
Gain(D, feature) &= Ent(D) - \sum_{v=1}^{2} \frac{|D^v|}{|D|} Ent(D^v) \\
Gain(D, A) &= 1 - (\frac{1}{2} \cdot 0.918 \cdot 2) = 0.082 \\
Gain(D, B) &= 1 - (\frac{1}{2} \cdot 1 \cdot 2) = 0 \\
Gain(D, C) &= 1 - (\frac{1}{2} \cdot 0.918 \cdot 2) = 0.082
\end{aligned}
\tag{2.1}
$$

由于 $Gain(D, A) = Gain(D, C)$, 任选其一作为划分属性。尝试选择属性 A。将 $A = 0$ 的训练样本划分为 D^1, $A = 1$ 的为 D^2 , 则

$$
\begin{aligned}
Ent(D^1) &= 0.918, Ent(D^2) = 0.928 \\
Gain(D^1, B) &= 0.918 - (\frac{1}{3} \cdot 0 + \frac{2}{3} \cdot 1) = 0.251 \\
Gain(D^1, C) &= Gain(D^2, B) = Gain(D^2, C) = 0.251
\end{aligned}
\tag{2.2}
$$

在根节点为属性 A 的情况下, 我们必须进一步划分节点才能够达到%100 的纯度. 但是, 若我们选择属性 C 作为划分节点, 将 $C = 0$ 的训练样本划分为 D^1, $C = 1$ 的为 D^2 , 则

$$
\begin{aligned}
Ent(D^1) &= 0.918, Ent(D^2) = 0.928 \\
Gain(D^1, A) &= Gain(D^2, A) = 0.918 - (\frac{1}{3} \cdot 0 + \frac{2}{3} \cdot 1) = 0.251 \\
Gain(D^1, B) &= Gain(D^2, B) = 0.928
\end{aligned}
\tag{2.3}
$$

可在只划分两次的基础上达到百分百的准确率。故选取 A 为根节点划分属性, B 为 D^1、D^2 的划分属性, 最终决策树如下。

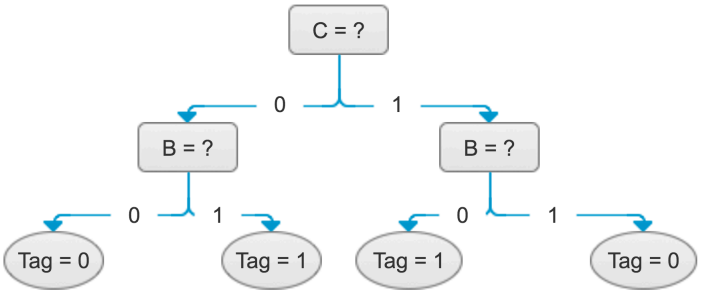

图 1: Decision Tree of Table 1

3 [40pts] Back Propagation

单隐层前馈神经网络的误差逆传播 (error BackPropagation，简称 BP) 算法是实际工程实践中非常重要的基础，也是理解神经网络的关键。

请编程实现 BP 算法, 算法流程如课本图 5.8 所示。详细编程题指南请参见链接：http://lamda.nju.edu.cn/ml2017/PS3/ML3_programming.html

在实现之后，你对 BP 算法有什么新的认识吗？请简要谈谈。

Solution. Sigmoid 的导数性质省去了很多计算, 且具有对阶跃函数良好的拟合, 计算 BP 的时候相对方便。

learn_rate 分别尝试设为 0.1, 0.5 和 0.8, epoch 分别尝试设为 100, 150, 200, 500, 最终在考虑时间消耗的情况下选取 learn_rate 为 0.5, epoch 为 200. 最终测试集上精度约为 93.7%

略微困惑的是, 将数据归一化步骤略去后，测试集上的准确度提高了 1.* 个百分点。

附加题 [30pts] Neural Network in Practice

在实际工程实现中，通常我们会使用已有的开源库，这样会减少搭建原有模块的时间。因此，请使用现有神经网络库，编程实现更复杂的神经网络。详细编程题指南请参见链接：http://lamda.nju.edu.cn/ml2017/PS3/ML3_programming.html

和上一题相比，模型性能有变化吗？如果有，你认为可能是什么原因。同时，在实践过程中你遇到了什么问题，是如何解决的？

Solution. 性能没有发生太大改变。